Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2

Abid Sarwar
Department of Computer Science & IT (Kishtwar Campus), University of Jammu, Jammu, India

Vinod Sharma
Department of Computer Science & IT, University of Jammu, Jammu, India

ABSTRACT

Artificial Intelligence is now a days gaining immense importance and is becoming a key technology in many fields ranging from banking industry, to travel industry, to communication industry, and to robotic industry. The use of Artificial Intelligence in medical diagnosis too is becoming increasingly common and has been used widely in the diagnosis of cancers, tumors, hepatitis, lung diseases, etc... The main aim of this paper is to build an Artificial Intelligent System that after analysis of certain parameters can predict that whether a person is diabetic or not. Diabetes is inability of body to manage the levels of sugar in the blood. It being one of the most chronic diseases around the world causes around 3.8 million deaths every year. Authors have identified 10 parameters that play an important role in diabetes and prepared a rich database of training data which served as the backbone of the prediction algorithm. Keeping in view this training data authors developed a system that uses the naïve-Bayes classification algorithm to serve the purpose. When the parameters of the test data are fed to the system, it anticipates & classifies the test data into one of the two categories viz diabetic & not diabetic. The performance of AI method when compared with the medical diagnosis system was found to be 95%. This system can be used to assist medical programs especially in geographically remote areas where expert human diagnosis not possible with an advantage of minimal expenses and faster results.

Keywords: Artificial Intelligence, Data Mining, Machine Learning, Diabetes, Naive Bayes classifier, Medical Diagnosis.

1. INTRODUCTION

Diabetes is a chronic condition that occurs when the body cannot produce enough or cannot effectively use insulin [1]. Insulin is a hormone produced by pancreas and is needed by our body to metabolize glucose. When the glucose level in body is not metabolized properly it keeps on circulating in the blood and causes damage to various tissues. Diabetes can mainly be of 3 types: Type-1 diabetes, Type-2 diabetes and Gestational diabetes. Type-1 diabetes results from non-production of insulin & Type-2 diabetes results from development of resistance of insulin, as a result of which the insulin produced is not able to metabolize the sugar levels properly. Gestational diabetes occurs in pregnant women, who develop a high blood glucose level during pregnancy who never had any previous such history. It may be preceded by development of type-2 diabetes. As reported by WHO & International Diabetic Federation in year 2010, the toll of diabetic patients was 285 million and this number is expected to grow to 483 million by 2030. WHO estimates that between 2010 to 2030 there will be an increase of 69% in adult diabetic population in developing countries and 20% increase of the same in developed countries. India having 50.8 million patients of this disease leads the world & is followed by China (43.2), United States (26.8). In 2010 diabetes caused 3.9 million deaths worldwide. The primary concern of AI in medicine is the construction of AI programs that can assert a medical doctor in performing expert diagnosis. These programs by making use of various computational sciences such as statistics and probability find out the hidden patterns from the training data and using these patterns they classify the test data into one the possible categories. The backbones of these AI programs are the various data sets prepared from various clinical cases which act as practical examples in training the system. The decision and recommendation prepared from these systems can be illustrated to the subjects after combining them with the experience of human expert.

2. METHODOLOGY

The authors selected the naïve Bayes classifier to train the system keeping in view its exceptional performance even in less amounts of training data. Naïve Bayes is considered to be one of the most efficient and effective inductive learning algorithms for machine learning and data mining. In 2006, Rich Caruana & Alexandru Niculescu-Mizil, did a comprehensive comparison of many classification algorithms and showed that naïve Bayes outperformed many of the counterparts such as boosted trees or random forests [4]. Its competitive performance is attributed to its principle of conditional independence assumption, which means that it assumes that the presence or absence of some parameters of a class to be independent to the presence or absence of some other parameters [3]. Thus each parameter has independent contribution to the prediction of the final result. Mathematically the probability model for a classifier is a conditional model $\mathbf{p(C \mid F_1, F_2 \ldots, F_n)}$, over a dependent class variable with a small number of outcomes or classes, conditional on several feature variables F1 to Fn. Using Bayes' theorem we rewrite the equation as :

$$\mathbf{p(C \mid F_1, F_2 \ldots, F_n) = p(C)\, p(F_1, F_2 \ldots, F_n) / p(F_1, F_2 \ldots, F_n)}$$

in simple English we write the above equation as :

Posterior = (Prior x Likelihood) / Evidence

Practically, we are interested only in the numerator part of this fraction, because the denominator is independent of C and the values of the features $\mathbf{F_i}$ are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model $\mathbf{p(C, F_1, F_2 \ldots, F_n)}$. This can be written as follows, using repeated applications of the definition of conditional probability

$$= \mathbf{p(C, F_1, F_2 \ldots, F_n)}$$

$$= \mathbf{p(C)\, p(F_1, F_2 \ldots, F_n \mid C)}$$

$$= \mathbf{p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1)\, p(F3, \ldots.. F_n \mid C, F_1, F_2)}$$

$$= \mathbf{p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1)\, p(F_3 \mid C, F_1, F_2)\, p(F_4 \ldots\ldots F_n \mid C, F_1, F_2, F_3)}$$

$$= \mathbf{p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1)\, p(F_3 \mid C, F_1, F_2) \ldots\ldots p(Fn \mid C, F_1, F_2, F_3 \ldots\ldots F_{n-1})}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature is conditionally independent of every other feature F_j for i not equal to j. This means that:

$$p(Fi \mid C, F_j) = p(Fi \mid C)$$

for i not equal to j and so the joint model can be expressed as

$$p(C, F_1, F_2 \ldots, F_n) = p(C)\, p(F_1 \mid C)\, p(F_2 \mid C)\, p(F_3 \mid C) \ldots..$$

$$= p(C) \prod p(F_i \mid C)$$

This means that under the above independence assumption the conditional distribution over the class variable C can be expressed like this:

$$p(C, F_1, F_2 \ldots, F_n) = 1/Z\, p(C) \prod p(Fi \mid C)$$

where Z (the evidence) is a scaling factor dependent only on $F_1, F_2, F_3 \ldots\ldots F_n$ **i.e.** a constant if the value of the feature variable s are known.

3. PREVIOUS WORK

It has been noted that the Machine learning algorithms are increasingly being used in solving problems in Medical Domains such as in Oncology [5,6,7,8], Urology[8,9], Hepatitis[9,13], Liver Pathology[10], Cardiology[15,16,17], Gynecology[18], Thyroid disorders [11,12], Tuberculosis[22], Neuropsychology[19], Perinatology [20] etc. Various algorithms have been used in different domains however the naïve Bayes algorithms have been noted to outperform most of the advanced and sophisticated algorithms in both medical diagnostic problems and in problems of non-medical domain [14]. Kononenko and others did a comprehensive comparison of naïve Bayes algorithm with six other algorithms and found that the naïve Bayes algorithm outperformed all other algorithms in 5 out of 8 problems of medical diagnostic domain. In a study, the Inductive Logic programming algorithms achieved a minimal classification accuracy of 12% to 29%, while the naive Bayes algorithm for similar problem achieved an accuracy of 35% [14]. Yet in another comparison between naïve Bayes and modern decision tree algorithms like C4.5 (Quinalan 1993) has proved that the naïve Bayes prediction capabilities are equally good as C4.5 (Langley, Iba, & Thomas 1992; Pazzani 1996; Kononenko 1990) [21].

4. PARAMETERS USED IN ESTIMATION

Since India is having the highest Diabetic population in the world so it was easy to collect the data about the patients who suffered from this disease. After a detailed study, authors identified ten best physiological parameters for the study which were so chosen that the values for them could be easily determined and could be assigned discrete values, for the sake of maintaining consistency. Table-I summaries the parameters chosen and their allowed values. A dataset of 415 cases was prepared by collecting the data randomly from different sections of the society with an aim to have a variety in the dataset. To maintain accuracy and to avoid errors, considerable care was taken to ensure that the database had correct values.

Table-I: Various parameters used and their allowed values

Parameter	Description	Allowed Values
Age	Age of the subject	Discrete Integer values
Family History	Whether any family member of the subject is suffering/ was suffering from diabetes.	Yes or No
Sex	Whether male or female	Male or Female
Smoking	Whether the subject does smoking or not.	Yes or No
Drinking	Whether the subject does drinking or not.	Yes or No
Fatigue	Does a person feel tired after doing a little work?	Yes or No
Thirst	Whether the subject frequently feels a strong desire to drink water. i.e how many times the subject drinks water.	Discrete Integer values
Frequency of Urination	How many times the subject passes urine in a day	Discrete Integer values
Height	Height of the subject	Discrete floating point values
Weight	Weight of the subject	Discrete floating point values

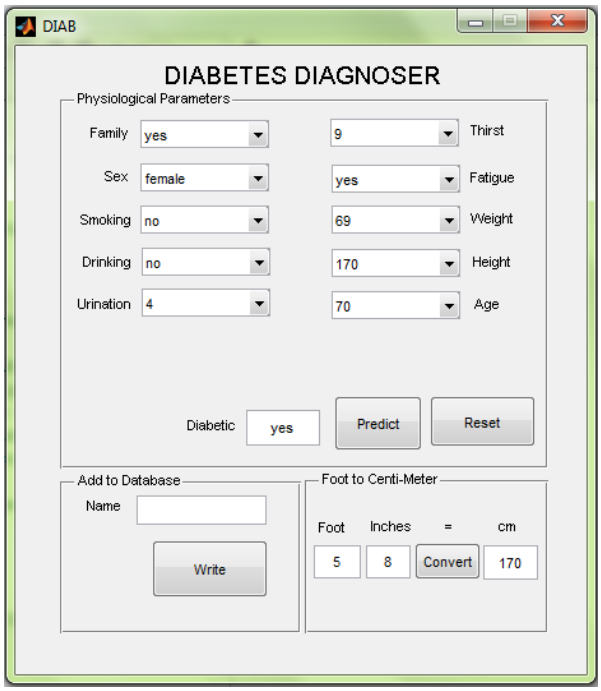
5. IMPLEMENTATION

As per the Conditional independence assumption of Bayes theorem, the presence or absence of some parameters of a class is independent to the presence or absence of some other parameters, making each parameter's contribution independent to the final result. The authors calculated the individual probability of all the variables for both Diabetic='Yes' & Diabetic = 'No'. For instance for a parameter "Frequency of Urination", the probability of both Diabetic ='Yes' & Diabetic = 'No' is calculated as:

P(Diabetic='Yes') given "Frequency of Urination" = 'Value from Test Data' & **P(Diabetic='No') given** "Frequency of Urination" = 'Value from Test Data'. In the similar way, the probabilities of all the parameters and stored there individual contribution to the final result in different variables can be calculated. To deal with the condition of zero probability values for some parameter, the authors made use of Laplace Correction. At last the consolidation to the contribution of all the individual variables, according to the test data gets classified into one of the two categories viz Diabetic or Not Diabetic. The development of the system is done using

MATLAB with SQL server as database. The experimental set up in execution is shown in the Figure-1.

Fig. 1: MATLAB Program in execution



6. CONCLUSION

This naïve Bayes classifier based system is very useful for diagnosis of diabetes. The reliability of the system was evaluated by computing the mean absolute error between the predicted values and exact values the cases. The results suggest that this system can perform good prediction with least error and finally this technique could be an important tool for supplementing the medical doctors in performing expert diagnosis. In this method the efficiency of Forecasting was found to be around 95%. Its performance can be further improved by identifying & incorporating various other parameters and increasing the size of training data.

7. ACKNOWLEDGEMENT

This paper has benefited by the discussions with people in the Area of Artificial Intelligence, Medical Sciences and Academicians up to a greater extent

8. REFERENCES

[1] Harris M, Zimmet P. Classification of diabetes mellitus and other categories of glucose intolerance. In Alberti K, Zimmet P, Defronzo R, editors. International Textbook of Diabetes Mellitus. Second Edition. Chichester: John Wiley and Sons Ltd; 1997. p9-23

[2] Williams textbook of endocrinology (12th ed.). Philadelphia: Elsevier/Saunders. pp. 1371–1435. ISBN 978-1437703245.

[3] Optimal naïve Bayes harry zahang

[4] An empirical comparison of supervised learning algos

[5] The use of artificial neural networks in decision support in cancer. Paulo J. Lisboa & Azzam F.G. Taktak

[6] Artificial Intelligence in Predicting Bladder Cancer Outcome: A Comparison of Neuro-Fuzzy Modeling and Artificial Neural Networks, James W. F. Catto, Derek A. Linkens, Maysam F. Abbod, Minyou Chen, Julian L. Burton, Kenneth M. Feeley, and Freddie C. Hamdy2

[7] Artificial neural networks for decision-making in urologic oncology, Anagnostou T, Remzi M, Lykourinas M, Djavan B

[8] Bratko I., Kononenko I, Learning Rules from Incomplete and Noisy Data, in B Phelps (ed.) Interactions in Arti_cial Intelligence and Statistical Methods, Hampshire, Technical Press.

[9] Experiments in automatic learning of medical diagnostic rules international school for synthesis of Expert knowledge Workshop, Bled, August 1981. Kononenko I, Bratko I, Roskar E.

[10] Lesmo. L. (et al) Learning of fuzzy production riles for medical diagnosis, In Gupta.M.M & Sanchez E.(eds), approximate reasoning in desion analysis, North-Holland 1982

[11] Hojker S, Kononenko I, Juka A, Fidler V. & Porenta. M, Expert System's Development in Management of Thyroid Disease, proc. European Congress for nuclear medicine, Milano, Sept., 1988

[12] Horn K.A, Compton P.Lazarus L., Quinlan J.R, an Expert System for Interpretation of Thyroid Assays in Clinical Laboratory, the Australian Computer Journal Vol. 17, No. 1, 1985, pp 7-11.

[13] Hepatitis disease diagnosis using Back-propagation and Naïve Bayes classifier.

[14] Machine learning for Medical Diagnosis: History, State of Art and Perspective; igor Knononkp……

[15] Bratko I., Mozetic I., Lavrac N., KARDIO: A study in deep and qualitative knowledge for expert system Cambridge,MA:MIT Press,1989.

[16] Catlett J., On changing continuous attributes into ordered discrete attributes, proc. European Working Session on Learning-91, Proto ,March 4-6 1991, pp. 164-178.

[17] Clark p . & Boswell R., Rule Induction with CN2: Some Recent Improvements, Proc. Europea Working Session on Learning-91,Porto, Portugal, March, 1991, pp.151-163.

[18] Nunez M., Decision Tree Induction Using Domain Knowledge, In: Wielinga B. et al. (eds.) Current Trends in Knowledge Acquisition, Amsterdam: IOS Press, 1990.

[19] Muggleton S., Inductive Acquisition of Knowledge, Turing Institute press & Addison-wesley, 1990.

[20] Kern j., Dezelic G., Tezak-Bencic M., Durrigl T., Medical Decision Making Using Inductive Learning program (in Croatian), Proc 1st Congress on Yougoslav Medical Informatics, Beogard, Dec 6-8, 1990, pp.221-228.

[21] The Optimality of Naïve Bayes, Harry Zhang. 2004, American Association for Artificial Intelligence.

[22] Comparative study of Naïve Bayes and KNN for Tuberculosis; Hardik Maniya , Mohsin Hasan, Komal Patil.